CRYSTAL STRUCTURE OF HUMAN α-GALACTOSIDASE

Related Applications

This application claims priority under 35 U.S.C. §119(e) from Provisional U.S. Patent Application Serial No. 60/536,451 filed January 13, 2004, and entitled CRYSTAL STRUCTURE OF HUMAN α-GALACTOSIDASE. The contents of the provisional application are hereby expressly incorporated by reference.

Government Support

Part of the work leading to the invention disclosed herein was made with Government support under Grant No. 1Z01AI000901 from NIAID. Accordingly, the U.S. Government has certain rights in this invention.

Field of the Invention

This invention relates to the X-ray crystal structure of the human α-galactosidase glycoprotein. More specifically, the invention relates to crystallized compositions of human α-galactosidase and to crystallized complexes of human α-galactosidase and its catalytic product α-galactose. The invention further relates to a computer programmed with the structure coordinates of the human α-galactosidase's active site wherein said computer is capable of displaying a three-dimensional representation of that active site. The invention also relates to methods for rational drug design based on the structural data for human α-galactosidase provided on computer readable media, as analyzed on a computer system having suitable computer algorithms.

Background of the Invention

The lysosomal enzyme α-galactosidase (α-GAL or α-Gal A, E.C. 3.2.1.22, SEQ ID NO:2) catalyzes the removal of galactose from oligosaccharides, glycoproteins, and glycolipids during the catabolism of macromolecules. Deficiencies in lysosomal enzymes lead to the accumulation of substrates in the tissues, conditions known as lysosomal storage diseases. In humans, the absence of functional α-GAL leads to the accumulation of galactosylated substrates (primarily globotriaosylceramide) in the tissues, leading to Fabry disease, an X-linked recessive disorder first described in 1898 (Fabry, J. *Arch. Dermatol. Syph.*, 1898, 43:187) characterized by chronic pain, ocular opacities, liver and kidney impairment, skin lesions, vascular deterioration and/or cardiac deficiencies (Brady, R. O., et al., *N. Engl. J. Med.*, 1967, 276:1163-7; Desnick, R. J., et al., In *The Metabolic and*

-2-

Molecular Bases of Inherited Disease 8th edit. -Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D., eds.-, 2001, pp. 3733-3774. McGraw-Hill, New York). Recombinant human α-GAL has the ability to restore enzyme function in patients (Schiffmann, R., et al., *JAMA* , 2001, 285:2743-9; Eng, C. M., et al., *N. Engl. J. Med.,* 2001, 345:9-16), and enzyme replacement therapy using α-GAL was recently approved in the United States as a treatment for Fabry disease. α-GAL became the second recombinant protein approved for the treatment of a lysosomal storage disorder (after β-glucosidase, a treatment for Gaucher disease - Beutler, E. & Grabowski, G. A., 2001, Gaucher Disease. In *The Metabolic and Molecular Bases of Inherited Disease* 8th edit. -Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D., eds.-McGraw-Hill, New York), and α-GAL represents one of a small number of recombinant human proteins approved for the treatment of any disease. A second treatment for Fabry disease (specific for the cardiac variant of the disease) uses galactose infusion, which presumably helps stabilize the mutant α-GAL protein (Frustaci, A., et al., *N. Engl. J. Med.,* 2001, 345:25-32). In addition to enzyme replacement therapy and galactose infusion, gene replacement therapy using the α-GAL gene shows potential as a treatment for Fabry disease (Park, J., et al., *Proc Natl Acad Sci U S A,* 2003, 100:3450-4).

There are currently two recombinant glycoprotein products, REPLAGAL™ (Transkaryotic Therapies, Inc., Cambridge, MA) and FABRAZYME™ (Genzyme, Inc., Cambridge, MA), available for enzyme replacement therapy used in the treatment of Fabry disease (Schiffmann, R., et al., *JAMA* , 2001, 285:2743-9; Eng, C. M., et al., *N. Engl. J. Med.,* 2001, 345:9-16). These two glycoproteins have identical amino acid sequences but are produced in different cell lines, resulting in different glycosylation at the N-linked carbohydrate attachment sites. REPLAGAL™ is produced in a genetically engineered human cell line, while FABRAZYME™ is produced in a Chinese hamster ovary (CHO) cell line. REPLAGAL™ contains a greater amount of complex carbohydrate while Fabrazyme contains a higher fraction of sialylated and phosphorylated carbohydrate (Lee, K., et al., *Glycobiology,* 2003, 13:305-13). Because the polypeptide sequence of the two glycoproteins is identical, these differences in carbohydrate composition are solely responsible for the differences in tissue distribution and dose response of the two enzyme replacement therapies.

α-GAL has also attracted attention for its ability to convert human blood group antigens. Recombinant α-GAL has been used to convert blood of type B into blood of type O, the universal donor type (Zhu, A., et al., *Arch. Biochem. Biophys.,* 1996, 327:324-9), a process currently in clinical trials.

-3-

Because of its utility in the treatment of Fabry disease and as a reagent for converting human blood types, much effort has been put into the expression and purification of large amounts of human α-GAL. The endogenous enzyme has been purified from human placenta (Mayes, J. S. & Beutler, E., *Biochim Biophys Acta,* 1977, 484:408-16), liver cells (Dean, K. J. & Sweeley, C. C., *J Biol Chem,* 1979, 254:9994-10000), spleen cells and plasma (Bishop, D. F. & Desnick, R. J., *J Biol Chem,* 1981, 256:1307-16), and fibroblasts (Lemansky, P., et al., *J Biol Chem,* 1987, 262:2062-5); recombinant enzyme has been produced in *E. coli* bacterial cells (Hantzopoulos, P. A. & Calhoun, D. H., *Gene,* 1987, 57:159-69), COS monkey cells (Tsuji, S., et al., *Eur J Biochem,* 1987, 165:275-80), CHO cells (Ioannou, Y. A., et al., *J Cell Biol,* 1992, 119:1137-50), baculovirus-infected Sf9 insect cells (Coppola, G., et a l., *Gene,* 1994, 144:197-203; Chen, Y., et al., *Protein Expr Purif,* 2000, 20:228-36), *Pichia pastoris* yeast cells (Chen, Y., et al., *Protein Expr Purif,* 2000, 20:472-84), transduced human bone marrow cells (Takenaka, T., et al., *Exp Hematol* 1999, 27:1149-59), and continuously cultured genetically engineered human fibroblasts (Schiffmann, R., et al., *JAMA ,* 2001, 285:2743-9). Despite the ability to successfully express and purify human α-GAL since 1977, the three-dimensional structure has not been solved, although a crystallization report appeared in 1994 (Murali, R., et al., *J. Mol. Biol.* 239:578-80). Structural analysis has been hindered by the heterogeneous carbohydrates on the glycoprotein, which comprise 5-15% of the mass of the secreted material and contain over 70 different species built upon 23 different core structures (Matsuura, F., et al., *Glycobiology* 1998, 8:329-39).

Thus, there is a great need to solve the crystal structure of α-GAL and, in particular, to delineate the active site of the enzyme. With this information, computer models of this active/binding site can be created and potential agonists and antagonists of α-GAL can be rationally designed.

Summary of the Invention

This invention provides the crystal structure of human α-GAL. The crystal structure has been solved by X-ray crystallography to a resolution of 3.25 Å. Based upon the crystal structure we have characterized human α-GAL in detail and identified the key amino acid residues that make up the active/binding site of the enzyme. These coordinates are useful in methods for designing agonists and antagonists of the enzyme, which in turn may be useful in treating Fabry and other diseases.

-4-

The invention also provides the X-ray structure coordinates of a complex comprising α-GAL and its catalytic product, α-galactose.

In another aspect the invention provides a computer programmed with the coordinates of the human α-GAL active/binding site, and with a program capable of converting those coordinates into a three-dimensional representation of the active site on a display connected to the computer.

In a further aspect, the invention provides a computer which, when programmed with at least a portion of the structural coordinates of human α-GAL and an X-ray diffraction data set of a different molecule or molecular complex, performs a Fourier transform of these structural coordinates of the human α-GAL coordinates and then processes the X-ray diffraction data into structure coordinates of the different molecule or molecular complex via the process of molecular replacement.

These and other objects of the invention will be described in further detail in connection with the detailed description of the invention.

Brief Description of the Sequences

SEQ ID NO:1 is the nucleotide sequence of the human α-GAL cDNA.

SEQ ID NO:2 is the predicted amino acid sequence of the translation product of human α-GAL cDNA (SEQ ID NO:1).

Brief Description of the Drawings

FIG. 1 (pp. 1-84 of 87 total pages of drawings) lists the atomic structure coordinates for human α-GAL as derived by X-ray diffraction from a crystal of human α-GAL dimer. The following abbreviations are used in FIG. 1: "Atom type" refers to the element whose coordinates are measured. The first letter in the column defines the element.

"X, Y, Z" crystallographically define the atomic position of the element measured.

"OCC" is an occupancy factor that refers to the fraction of the molecules in which each atom occupies the position specified by the coordinates. A value of "1" indicates that each atom has the same conformation, i.e., the same position, in all molecules of the crystal.

"B" is a thermal factor that measures movement of the atom around its atomic center.

FIG. 2 shows a diagram of a computer used to generate a three-dimensional graphical representation of a molecule or molecular complex according to this invention.

FIG. 3 shows a cross section of a magnetic storage medium.

-5-

FIG. 4 shows a cross section of an optically-readable data storage medium.

FIG. 5 is a schematic representation of the human α-GAL active site with a galactose molecule.

Detailed Description of the Invention

As mentioned above, we have solved the three-dimensional X-ray crystal structure of human α-galactosidase. The atomic coordinate data is presented in FIG. 1.

In order to use the structure coordinates generated for the human α-galactosidase, its active site or portions or homologues thereof, it is often times necessary to convert them into a three-dimensional shape. This is achieved through the use of commercially available software that is capable of generating three-dimensional graphical representations of molecules or portions thereof from a set of structure coordinates.

An "active site", also referred to as "binding site" elsewhere herein, is of significant utility in fields such as drug discovery. The association of natural ligands or substrates with the active site(s) (or "binding pocket") of their corresponding receptors or enzymes is the basis of many biological mechanisms of action. Similarly, many drugs exert their biological effects through association with the binding pockets of receptors and enzymes. Such associations may occur with all or any parts of the binding pocket. An understanding of such associations will help lead to the design of drugs having more favorable associations with their target receptor or enzyme, and thus, improved biological effects. Therefore, this information is valuable in designing potential agonists and antagonists of the binding sites of biologically important targets.

The term "active site" (or "binding pocket"), as used herein, refers to a specific region of an enzyme, that, as a result of its shape, favorably associates with its substrate and catalysis occurs.

We have identified at least one active site per monomer in human α-GAL, which is a good target for designing agonists and/or antagonists and/or inhibitors.

The terms "α-GAL-like binding pocket", as used herein, refers to a portion of a molecule or molecular complex whose shape is sufficiently similar to the human α-GAL binding pocket, so as to bind common ligands. These commonalties of shape are defined by a root mean square deviation from the structure coordinates of the backbone atoms of the amino acids that make up these binding pockets in the human α-GAL structure (as set forth

-6-

in FIG. 1) of not more than 1.5 Å. The method of performing this calculation is described below.

The x-ray structure reveals human α-GAL as a homodimeric glycoprotein with each monomer composed of two domains, a $(\beta/\alpha)_8$ domain containing the active site and a C-terminal domain containing eight antiparallel β strands on two sheets in a β sandwich. After removal of the 31 residue signal sequence, the first domain extends from residues 32 to 330 and contains the active site formed by the C-terminal ends of the β strands at the center of barrel, a typical location for the active site in $(\beta/\alpha)_8$ domains. The second domain, comprised of residues 331 to 429, packs against the first with an extensive interface, burying 2500 $Å^2$ of surface area within one monomer. The dimer has overall protein dimensions of approximately 75 x 75 x 50Å. The molecule is concave in the third dimension and varies in thickness from approximately 20 to 50Å. Electron density is visible for 390 and 391 amino acid residues (out of 398 total) in the two copies of the monomer in the crystallographic asymmetric unit; the missing residues occur at the C-terminus. The two monomers pack with an interface that extends the 75Å width of the dimer and buries 2200 $Å^2$ of surface area. In the dimer interface, 30 residues from each monomer contribute to the interface, from loops β1-α1, β6-α6, β7-α7, β8-α8, β11-β12, and β15-β16. The dimer is markedly negatively charged, as seen in a surface electrostatic potential. With 47 carboxylate groups and only 36 basic residues in the 398 residues in the molecule, the overall charge per monomer is expected to be −11 at neutral pH. The carboxylates are most concentrated around the active site, but in the low pH of the lysosome, many of these groups become protonated, reducing the charge on the molecule. In addition to the negative charges on the protein, the N-linked carbohydrate is highly phosphorylated and sialylated (Lee, K., et al., *Glycobiology,* 2003, 13:305-13) (see below), further increasing its negative electrostatic potential. The N-linked carbohydrates fall distal to the active sites. Each monomer contains the three N-linked carbohydrate sites, five disulfide bonds (C52-C94, C56-C63, C142-C172, C202-C223, and C378-C382), two unpaired cysteines (C90 and C174), and three *cis* prolines (P210, P380, and P389).

As mentioned above, the C-terminal seven and eight residues of each chain have no electron density associated with them and are presumably disordered. This disorder is consistent with the observation of slight heterogeneity in the C-terminus of recombinant human α-GAL, where the truncation of one or two residues from the C-terminus can occur but has no effect upon the activity of the enzyme (Lee, K., et al., *Glycobiology,* 2003, 13:305-

-7-

13). The structure offers no support for the observation that the removal of 2 to 10 residues from the C-terminus increases the activity of α-GAL (Miyamura, N., et al., *J Clin Invest*, 1996, 98:1809-17), because the final residue seen in the structure falls at least 45Å from each active site and on the opposite face of the molecule.

In both the native and galactose-soaked crystal structures, electron density appears in the two crystallographically-independent active sites. In the galactose-soaked crystal, this density represents α-galactose, the normal catalytic product of the enzyme (K_i ~1mM). In the native structure, this density most likely derives from the cryoprotectant ethylene glycol, a weak inhibitor of glycoside hydrolases (Tsitsanou, K. E., et al., *Protein Sci*, 1999, 8:741-9), analogous to the insertion of glycerol into carbohydrate binding sites on proteins (Garman, S. C., et al., *Structure*, 2002, 10:425-434; Tsitsanou, K. E., et al., *Protein Sci*, 1999, 8:741-9; Schmidt, A., et al., *Protein Sci*, 1998, 7:2081-8). The two active sites of the dimer are separated by approximately 50 Å. As the enzyme shows little change between the liganded and unliganded structures, there is no evidence for cooperativity between the two sites, although the biochemical evidence is mixed (Dean, K. J. & Sweeley, C. C., *J Biol Chem*, 1979, 254:9994-10000; Bishop, D. F. & Desnick, R. J., *J Biol Chem*, 1981, 256:1307-16).

We have determined that human α-GAL binds α-galactose by making specific contacts to each functional group on the monosaccharide. Residues from seven loops in domain 1 form the *active site*: β1-α1, β2-α2, β3-α3, β4-α4, β5-α5, β6-α6, and β7-α7. The active site is formed by the side chains of residues W47, D92, D93, Y134, C142, K168, D170, E203, L206, Y207, R227, D231, D266, and M267 (a.k.a. Trp47, Asp92, Asp93, Tyr134, Cys142, Lys168, Asp170, Glu203, Leu206, Tyr207, Arg227, Asp231, Asp266, and Met267). Thus, a binding pocket defined by the structural coordinates of these amino acids, as set forth in FIG. 1; or a binding pocket whose root mean square deviation from the structure coordinates of the backbone atoms of these amino acids is not more than 1.5 Å is considered a human α-GAL-like binding pocket of this invention. In important embodiments, C172 (Cys172) makes a disulfide bond to C142 (Cys142).

In the α-GAL/α-NAGAL family, specificity for the 2 position on the galactose ligand occurs via the β5-α5 loop. This was called the "N-acetyl recognition loop" in α-NAGAL (Garman, S. C., et al., *Structure*, 2002, 10:425-434); in the overall α-GAL/α-NAGAL family "2 position recognition loop" or "2 loop" is appropriate. This loop falls near the boundary of exons 4 and 5 of animal α-GAL/α-NAGAL, which have a small insertion in this region, resulting in a short helical stretch at the top of the β5 strand; this insertion is absent in other

-8-

species. Plant and fungal α-GALs use a Cys and a Trp on this loop to coordinate the 2-hydroxyl on galactose; animal α-GAL uses a Glu and a Leu to recognize the 2-hydroxyl while animal α-NAGAL uses a Ser and an Ala to recognize an N-acetyl at the 2 position. In the animal enzymes, the larger Glu and Leu side chains sterically block the larger N-acetyl substituent, while the smaller Ser and Ala side chains nicely accommodate an N-acetyl group and tolerate a hydroxyl group.

With three different conformations in the 2 loop now identified, the substrate specificity of the other members of the family can be categorized by homology. For example, genome sequencing of *Drosophila melanogaster* and *Anopheles gambiae* have each identified pairs of genes in the α-GAL family. By examination of the sequences in the 2 loop, two are clearly α-NAGALs while the other two appear to be α-GALs. Surprisingly, *Aspergillus niger* contains an enzyme identified as α-GAL that, although only 30% identical to the animal protein sequences, contains a 2 loop virtually identical to animal α-NAGALs We predict this enzyme is primarily an α-NAGAL with partial α-GAL activity, much like human α-NAGAL, which was originally thought to be an α-GAL based upon similar activity (Dean, K. J., et al., *Biochem. Biophys. Res. Commun.*, 1977, 77:1411-7; Schram, A. W., et al., *Biochim. Biophys. Acta*, 1977, 482:138-44).

Although human α-GAL makes contacts to each functional group on the α-galactose ligand, the enzyme shows little specificity for the distal portion of the substrate beyond the glycosidic linkage, and the active site cleft is found in a broad opening on the concave surface of the enzyme. The lack of substrate specificity of human α-GAL beyond the terminal α-galactose differs slightly from the specificity of other α-GALs, which act only upon substrates containing terminal α1-6 galactose groups (Kim,W.D., et al., *Phytochemistry*, 2002, 61:621-30). This increased specificity of plant α-GALs may derive from their monomeric structure, as residues buried in the dimer interface of animal α-GALs (*e.g.*, those on the β1-α1 loop - Fujimoto, Z., et al., *J Biol Chem*, 2003, 278:20313-8) are available for ligand recognition in monomeric α-GALs.

Both α-GALs and α-NAGALs are α retaining exoglycosidases, where both the substrate and product of the catalytic reactions are α anomers at the 1 position on the galactose ring. This retention of anomeric configuration is accomplished by a double displacement catalytic mechanism where the anomeric carbon undergoes two successive nucleophilic attacks (Vasella, A., et al., *Curr Opin Chem Biol*, 2002, 6:619-29). The two

-9-

sequential inversions of the anomeric carbon lead to retention of the configuration at the end of the catalytic cycle. In two α-GALs from different species, peptic digestion of covalently trapped intermediates has identified the specific aspartic acid acting as the catalytic nucleophile (Hart, D. O., et al., *Biochemistry,* 2000, 39:9826-36; Ly, H. D., et al., *Carbohydr. Res.,* 2000, 329:539-47). These data, combined with the high resolution structure of chicken α-NAGAL, predict the catalytic mechanism of human α-GAL. In human α-GAL, the first nucleophilic attack upon the substrate comes from D170, cleaving the glycosidic linkage and leading to a covalent enzyme-intermediate complex. In the second step of the reaction, a water molecule (deprotonated by D231) attacks C1 of the covalent intermediate, liberating the second half of the catalytic product and regenerating the enzyme in its initial state. Human α-GAL operates most efficiently at low pH, consistent with its highly acidic composition and its lysosomal location.

Retaining glycosidases typically have distances of 5–6Å between catalytic carboxylates, while inverting glycosidases typically have distances of 9–11Å between these residues (McCarter, J. D. & Withers, S. G., *Curr. Opin. Struct. Biol.* 1994, 4:885-92). From these distances, it has been possible to reliably predict the mechanism and function of a glycosidase given its structure. However, this rule must be reconsidered in light of the new structures in the α-GAL/α-NAGAL family: for the known structures in the family, the closest approach of the two catalytic carboxylates is 6.5-7Å, among the largest distances seen for retaining glycosidases.

It will be readily apparent to those of skill in the art that the numbering of amino acids in other isoforms of human α-GAL may be different than that set forth for herein. Corresponding amino acids in other isoforms of human α-GAL are easily identified by visual inspection of the amino acid sequences or by using commercially available homology software programs. Each of those amino acids of human α-GAL is defined by a set of structure coordinates set forth in FIG. 1. The term "structure coordinates" refers to Cartesian coordinates derived from mathematical equations related to the patterns obtained on diffraction of a monochromatic beam of X-rays by the atoms (scattering centers) of a protein or protein-ligand complex in crystal form. The diffraction data are used to calculate an electron density map of the repeating unit of the crystal. The electron density maps are then used to establish the positions of the individual atoms of the enzyme or enzyme complex.

Those of skill in the art understand that a set of structure coordinates for an enzyme or an enzyme-complex or a portion thereof, is a relative set of points that define a shape in three

dimensions. Thus, it is possible that an entirely different set of coordinates could define a similar or identical shape. Moreover, slight variations in the individual coordinates will have little effect on overall shape. In terms of binding pockets, these variations would not be expected to significantly alter the nature of ligands that could associate with those pockets.

5   The term "associating with" refers to a condition of proximity between a chemical entity or compound, or portions thereof, and a binding pocket or binding site on a protein. The association may be non-covalent--wherein the juxtaposition is energetically favored by hydrogen bonding or van der Waals or electrostatic interactions--or it may be covalent.

        The variations in coordinates discussed above may be generated because of
10   mathematical manipulations of the human α-GAL structure coordinates. For example, the structure coordinates set forth in FIG. 1 could be manipulated by crystallographic permutations of the structure coordinates, fractionalization of the structure coordinates, integer additions or subtractions to sets of the structure coordinates, inversion of the structure coordinates or any combination of the above.

15   Alternatively, modifications in the crystal structure due to mutations, additions, substitutions, and/or deletions of amino acids, or other changes in any of the components that make up the crystal could also account for variations in structure coordinates. If such variations are within an acceptable standard error as compared to the original coordinates, the resulting three-dimensional shape is considered to be the same. Thus, for example, a ligand
20   (e.g., substrate) that bound to the α-GAL active site would also be expected to bind to another binding pocket whose structure coordinates defined a shape that fell within the acceptable error.

        Various computational analyses are therefore necessary to determine whether a molecule or the binding pocket portion thereof is sufficiently similar to the α-GAL
25   active/binding site described above. Such analyses may be carried out in well known software applications, such as the Molecular Similarity application of QUANTA™ (Molecular Simulations Inc., San Diego, CA.) version 4.1, and as described in the accompanying User's Guide.

        The Molecular Similarity application permits comparisons between different
30   structures, different conformations of the same structure, and different parts of the same structure. The procedure used in Molecular Similarity to compare structures is divided into four steps: 1) load the structures to be compared; 2) define the atom equivalences in these structures; 3) perform a fitting operation; and 4) analyze the results.

-11-

Each structure is identified by a name. One structure is identified as the target (i.e., the fixed structure); all remaining structures are working structures (i.e., moving structures). Since atom equivalency within QUANTA™ is defined by user input, for the purpose of this invention we will define equivalent atoms as protein backbone atoms (N, Cα, C and O) for all conserved r esidues b etween t he t wo s tructures b eing c ompared. We also consider o nly rigid fitting operations.

When a rigid fitting method is used, the working structure is translated and rotated to obtain an optimum fit with the target structure. The fitting operation uses an algorithm that computes the optimum translation and rotation to be applied to the moving structure, such that the root mean square difference of the fit over the specified pairs of equivalent atom is an absolute minimum. This number, given in angstroms (Å), is reported by QUANTA™.

For the purpose of this invention, any molecule or molecular complex or binding pocket thereof that has a root mean square deviation of conserved residue backbone atoms (N, Cα, C and O) of less than 1.5 Å when superimposed on the relevant backbone atoms described by structure coordinates listed in FIG. 1 are considered identical. More preferably, the root mean square deviation is less than 1.0 Å.

The term "root mean square deviation" means the square root of the arithmetic mean of the squares of the deviations from the mean. It is a way to express the deviation or variation from a trend or object. For purposes of this invention, the "root mean square deviation" defines the variation in the backbone of a protein from the backbone of human α-GAL or a binding pocket portion thereof, as defined by the structure coordinates of human α-GAL described herein.

Therefore, according to one aspect of the invention a computer is provided for producing:

(a) a three-dimensional representation of a molecule or molecular complex, wherein said molecule or molecular complex comprises a binding pocket defined by structure coordinates of human α-galactosidase amino acids W47, D92, D93, Y134, C142, K168, D170, E203, L206, Y207, R227, D231, D266, and M267, according to FIG. 1; or

b) a three-dimensional representation of a homologue of said molecule or molecular complex, wherein said homologue comprises a binding pocket that has a root mean square deviation from the backbone atoms of said amino acids of not more than 1.5 Å, wherein said computer comprises:

-12-

(i)   a machine-readable data storage medium comprising a data storage material encoded with machine-readable data, wherein said data comprises the structure coordinates of human α-galactosidase amino acids W47, D92, D93, Y134, C142, K168, D170, E203, L206, Y207, R227, D231, D266, and M267, according to FIG. 1;

(ii)   a working memory for storing instructions for processing said machine-readable data;

(iii) a central-processing unit coupled to said working memory and to said machine-readable data storage medium for processing said machine readable data into said three-dimensional representation; and

(iv)   a display coupled to said central-processing unit for displaying said three-dimensional representation.

In an important embodiment, C172 makes a disulfide bond to C142.

According to another aspect of the invention, a computer for producing a three-dimensional representation of a molecule or molecular complex defined by structure coordinates of all of the human α-GAL amino acids set forth in FIG. 1, or a three-dimensional representation of a homologue of said molecule or molecular complex, is provided. The homologue comprises a binding pocket that has a root mean square deviation from the backbone atoms of said amino acids of not more than 1.5 Å. In this aspect of the invention, a machine readable data contains the coordinates of all of human α-GAL.

According to a further aspect, the invention provides a computer for determining at least a portion of the structure coordinates corresponding to X-ray diffraction data obtained from a molecule or molecular complex, wherein said computer comprises:

(a) a machine-readable data storage medium comprising a data storage material encoded with machine-readable data, wherein said data comprises at least a portion of the structural coordinates of human α-GAL according to FIG. 1;

(b) a machine-readable data storage medium comprising a data storage material encoded with machine-readable data, wherein said data comprises X-ray diffraction data from said molecule or molecular complex;

(c) a working memory for storing instructions for processing said machine-readable data of (a) and (b);

-13-

(d) a central-processing unit coupled to said working memory and to said machine-readable data storage medium of (a) and (b) for performing a Fourier transform of the machine readable data of (a) and for processing said machine readable data of (b) into structure coordinates; and

5   (e) a display coupled to said central-processing unit for displaying said structure coordinates of said molecule or molecular complex.

FIG. 2 demonstrates one version of the foregoing aspects. System 10 includes a computer 11 comprising a central processing unit ("CPU") 20, a working memory 22 which may be, e.g., RAM (random-access memory) or "core" memory, mass storage memory 24
10   (such as one or more disk drives or CD-ROM drives), one or more cathode-ray tube ("CRT") display terminals 26, one or more keyboards 28, one or more input lines 30, and one or more output lines 40, all of which are interconnected by a conventional bi-directional system bus 50.

Input hardware 36, coupled to computer 11 by input lines 30, may be implemented in
15   a variety of ways. Machine-readable data of this invention may be inputted via the use of a modem or modems 32 connected by a telephone line or dedicated data line 34. Alternatively or additionally, the input hardware 36 may comprise CD-ROM drives or disk drives 24. In conjunction with display terminal 26, keyboard 28 may also be used as an input device.

Output hardware 46, coupled to computer 11 by output lines 40, may similarly be
20   implemented by conventional devices. By way of example, output hardware 46 may include CRT display terminal 26 for displaying a graphical representation of a binding pocket of this invention using a program such as QUANTA™ as described herein. Output hardware might also include a printer 42, so that hard copy output may be produced, or a disk drive 24, to store system output for later use.

25   In operation, CPU 20 coordinates the use of the various input and output devices 36, 46, coordinates data accesses from mass storage 24 and accesses to and from working memory 22, and determines the sequence of data processing steps. A number of programs may be used to process the machine-readable data of this invention. Such programs are discussed in reference to the computational methods of drug discovery as described herein.
30   Specific references to components of the hardware system 10 are included as appropriate throughout the following description of the data storage medium.

FIG. 3 shows a cross section of a magnetic data storage medium 100 which can be encoded with a machine-readable data that can be carried out by a system such as system 10 of FIG. 2. Medium 100 can be a conventional floppy diskette or hard disk, having a suitable

-14-

substrate 101, which may be conventional, and a suitable coating 102, which may be conventional, on one or both sides, containing magnetic domains (not visible) whose polarity or orientation can be altered magnetically. Medium 100 may also have an opening (not shown) for receiving the spindle of a disk drive or other data storage device 24.

5   The magnetic domains of coating 102 of medium 100 are polarized or oriented so as to encode in manner which may be conventional, machine readable data such as that described herein, for execution by a system such as system 10 of FIG. 2.

        FIG. 4 shows a cross section of an optically-readable data storage medium 110 which also can be encoded with such a machine-readable data, or set of instructions, which can be

10   carried out by a system such as system 10 of FIG. 2. Medium 110 can be a conventional compact disk read only memory (CD-ROM) or a rewritable medium such as a magneto-optical disk which is optically readable and magneto-optically writable. Medium 100 preferably has a suitable substrate 111, which may be conventional, and a suitable coating 112, which may be conventional, usually of one side of substrate 111.

15   In the case of CD-ROM, as is well known, coating 112 is reflective and is impressed with a plurality of pits 113 to encode the machine-readable data. The arrangement of pits is read by reflecting laser light off the surface of coating 112. A protective coating 114, which preferably is substantially transparent, is provided on top of coating 112.

        In the case of a magneto-optical disk, as is well known, coating 112 has no pits 113,

20   but has a plurality of magnetic domains whose polarity or orientation can be changed magnetically when heated above a certain temperature, as by a laser (not shown). The orientation of the domains can be read by measuring the polarization of laser light reflected from coating 112. The arrangement of the domains encodes the data as described above.

        Thus, in accordance with the present invention, X-ray coordinate data capable of

25   being processed into a three dimensional graphical display of a molecule or molecular complex which comprises an α-GAL-like binding pocket is stored in a machine-readable storage medium.

        The human α-GAL X-ray coordinate data, when used in conjunction with a computer programmed with software to translate those coordinates into the 3-dimensional structure of a

30   molecule or molecular complex comprising an α-GAL-like binding pocket may be used for a variety of purposes, such as drug discovery.

        For example, the structure encoded by the data may be computationally evaluated for its ability to associate with chemical entities. Chemical entities that associate with human α-

-15-

GAL may inhibit that enzyme, and are potential drug candidates. Alternatively, the structure encoded by the data may be displayed in a graphical three-dimensional representation on a computer screen. This allows visual inspection of the structure, as well as visual inspection of the structure's association with chemical entities.

Thus, according to another aspect the invention relates to a method for evaluating the potential of a chemical entity to associate with:

a) a molecule or molecular complex comprising a binding pocket defined by structure coordinates of human α-galactosidase amino acids W47, D92, D93, Y134, C142, K168, D170, E203, L206, Y207, R227, D231, D266, and M267, according to FIG. 1, or

b) a homologue of said molecule or molecular complex, wherein said homologue comprises a binding pocket that has a root mean square deviation from the backbone atoms of said amino acids of not more than 1.5 Å. The method comprises the steps of:

i) employing computational means to perform a fitting operation between the chemical entity and a binding pocket of the molecule or molecular complex; and

ii) analyzing the results of said fitting operation to quantify the association between the chemical entity and the binding pocket.

The term "chemical entity," as used herein, refers to chemical compounds, complexes of at least two chemical compounds, and fragments of such compounds or complexes.

Alternatively, the structural coordinates of the human α-GAL binding pocket can be utilized in a method for identifying a potential agonist or antagonist of a molecule comprising a human α-GAL-like binding pocket. The method comprises the steps of:

a) using the atomic coordinates of human α-galactosidase amino acids W47, D92, D93, Y134, C142, K168, D170, E203, L206, Y207, R227, D231, D266, and M267, according to FIG. 1 ± a root mean square deviation from the backbone atoms of said amino acids of not more than 1.5 Å, to generate a three-dimensional structure of molecule comprising α-GAL-like binding pocket;

b) employing said three-dimensional structure to design or select said potential agonist or antagonist;

c) synthesizing said agonist or antagonist; and

d) contacting said agonist or antagonist with said molecule to determine the ability of said potential agonist or antagonist to interact with said molecule.

In important embodiments, the atomic coordinates of all the amino acids of human α-GAL according to FIG. 1 ± a root mean square deviation from the backbone atoms of said

-16-

amino acids of not more than 1.5 Å, are used to generate a three-dimensional structure of molecule comprising an α-GAL-like binding pocket.

For the first time, the present invention permits the use of molecular design techniques to identify, select and design chemical entities, including agonists and antagonists, capable of binding to human α-GAL-like binding pockets. Because of the present invention, the necessary information for designing new chemical entities and compounds that may interact with human α-GAL-like binding pockets, in whole or in part, is provided.

Throughout this section, discussions about the ability of an entity to bind to, associate with or inhibit a human α-GAL-like binding pocket refers to features of the entity alone. Assays to determine if a compound binds to human α-GAL are well known in the art and are exemplified below.

The design of compounds that bind to or inhibit human α-GAL-like binding pockets according to this invention generally involves consideration of two factors. First, the entity must be capable of physically and structurally associating with parts or all of the human α-GAL -like binding pockets. Non-covalent molecular interactions important in this association include hydrogen bonding, van der Waals interactions, hydrophobic interactions and electrostatic interactions.

Second, the entity must be able to assume a conformation that allows it to associate with the human α-GAL-like binding pocket directly. Although certain portions of the entity will not directly participate in these associations, those portions of the entity may still influence the overall conformation of the molecule. This, in turn, may have a significant impact on potency. Such conformational requirements include the overall three-dimensional structure and orientation of the chemical entity in relation to all or a portion of the binding pocket, or the spacing between functional groups of an entity comprising several chemical entities that directly interact with the human α-GAL-like binding pocket or homologues thereof.

The potential inhibitory or binding effect of a chemical entity on a human α-GAL-like binding pocket may be analyzed prior to its actual synthesis and testing by the use of computer modeling techniques. If the theoretical structure of the given entity suggests insufficient interaction and association between it and the human α-GAL-like binding pocket, testing of the entity is obviated. However, if computer modeling indicates a strong interaction, the molecule may then be synthesized and tested for its ability to bind to a human α-GAL-like binding pocket. This may be achieved by testing the ability of the molecule to

-17-

inhibit human α-GAL using assays described in the art. In this manner, synthesis of inoperative compounds may be avoided.

A potential inhibitor of a human α-GAL-like binding pocket may be computationally evaluated by means of a series of steps in which chemical entities or fragments are screened and selected for their ability to associate with the human α-GAL-like binding pockets.

One skilled in the art may use one of several methods to screen chemical entities or fragments for their ability to associate with a human α-GAL-like binding pocket. This process may begin by visual inspection of, for example, a human α-GAL-like binding pocket on the computer screen based on the human α-GAL structure coordinates in FIG. 1 or other coordinates which define a similar shape generated from the machine-readable storage medium. Selected fragments or chemical entities may then be positioned in a variety of orientations, or docked, within that binding pocket as defined supra. Docking may be accomplished using software such as QUANTA™ and Sybyl™, followed by energy minimization and molecular dynamics with standard molecular mechanics force fields, such as Charmm™ and Amber™.

Specialized computer programs may also assist in the process of selecting fragments or chemical entities. These include: GRID (P. J. Goodford, *J. Med. Chem.*, 1985, 28:849-857), available from Oxford University, Oxford, UK; MCSS (A. Miranker et al., *Proteins: Structure, Function and Genetics,* 1991, 11:29-34), available from Molecular Simulations, San Diego, CA; AUTODOCK (D. S. Goodsell et al., *Proteins: Structure. Function, and Genetics*, 1990, 8:195-20), available from Scripps Research Institute, La Jolla, CA; DOCK (I. D. Kuntz et al., *J. Mol. Biol.,* 1982, 161:269-288), available from University of California, San Francisco, CA.

Other suitable software that can be used to view, analyze, design, and/or model a protein, and/or protein fragments, include but are not limited to: ALCHEMY™, LABVISION™, SYBYL™, MOLCADD™, LEAPFROG™, MATCHMAKER™, GENEFOLD™ and SITEL™ (available from Tripos Inc., St. Louis, MO); QUANTA™, CERIUS2™, X-PLOR™, CNS™, CATALYST™, MODELLER™, CHEMX™, LUDI™, INSIGHT™, DISCOVER™, CAMELEON™ and IDITIS™ (available from Accelrys Inc., Princeton N.J.); RASMOL™ (available from Glaxo Research and Development, Greenford, Middlesex, U.K.); MOE™ (available from Chemical Computing Group, Montreal, Quebec, Canada); MAESTRO™ (available from Shrodinger Inc.,); MIDAS/MIDASPLUS™ (available from UCSF, San Francisco, CA); VRML (webviewer--freeware on the internet);

-18-

CHIME (MDL--freeware on the internet); MOIL (available from University of Illinois, Urbana-Champaign, IL); MACROMODEL™ and GRASP™ (available from Columbia University, New York, NY); RIBBON™ (available from University of Alabama, Tuscaloosa, AL); NAOMI™ (available from Oxford University, Oxford, UK); EXPLORER EYECHEM™ (available from Silicon Graphics Inc., Mountain View, CA); UNIVISION™ (available from Cray Research Inc., Seattle, WA); MOLSCRIPT™ and O (available from Uppsala University, Uppsala, Sweden); CHEM 3D™ and PROTEIN EXPERT™ (available from Cambridge Scientific, MA); CHAIN™ (available from Baylor College of Medicine, Houston, TX); SPARTAN™, MACSPARTAN™ and TITAN™ (available from Wavefunction Inc., Irvine, CA); VMD™ (available from U. Illinois/Beckman Institute); SCULPT™ (available from Interactive Simulations, Inc., Portland, OR); PROCHECK™ (available from Brookhaven National Laboratory, Upton, NY); DGEOM (available from QCPE--Quantum Chemistry Program Exchange, Indiana University Bloomington, IN); RE_VIEW (available from Brunel University, London, UK); XmoI (available from Minnesota Supercomputing Center, University of Minnesota, Minneapolis, MN); HYPERCHEM™ (available from Hypercube, Inc., Gainesville, FL); MD Display (available from University of Washington, Seattle, WA.); PKB (available from National Center for Biotechnology Information, NIH, Bethesda, MD); Molecular Discovery Programmes (available from Molecular Discovery Limited, Mayfair, London); GROWMOL™ (available from Thistlesoft, Morris Township, N.J.); MICE (available from The San Diego Supercomputer Center. La Jolla, CA); Yummie and MCPro (available from Yale University, New Haven, CT); CAVEAT™ (P. A. Bartlett et al, In *"Molecular Recognition in Chemical and Biological Problems"*, Special Pub., Royal Chem. Soc., 1989, 78:182-196; G. Lauri and P. A. Bartlett, *J. Comput. Aided Mol. Des.*, 1994, 8:51-66), available from the University of California, Berkeley, CA; 3D Database systems such as ISIS™ (MDL Information Systems, San Leandro, CA). This area is reviewed in Y. C. Martin, "3D Database Searching in Drug Design", *J. Med. Chem.*, 1992, 35:2145-2154; HOOK™ (M. B. Eisen et al, *Proteins: Struct., Funct., Genet.*, 1994, 19:199-221), available from Molecular Simulations, San Diego, CA; and upgraded versions thereof.

Once suitable chemical entities or fragments have been selected, they can be assembled into a single compound or complex. Assembly may be preceded by visual inspection of the relationship of the fragments to each other on the three-dimensional image displayed on a computer screen in relation to the structure coordinates of human α-GAL.

-19-

This would be followed by manual model building using software such as QUANTA™ or SYBYL™.

Instead of proceeding to build an inhibitor of human α-GAL-like binding pocket in a step-wise fashion one fragment or chemical entity at a time as described above, inhibitory or other human α-GAL binding compounds may be designed as a whole or "de novo" using either an empty binding site or optionally including some portion(s) of a known inhibitor(s).

Other molecular modeling techniques may also be employed in accordance with this invention [see, e.g., N. C. Cohen et al., *J. Med. Chem.*, 1990, 33:883-894; see also, M. A. Navia and M. A. Murcko, *Curr. Opin. in Struct. Biology*, 1992, 2:202-210; L. M. Balbes et al., "A Perspective of Modern Methods in Computer-Aided Drug Design", in *Reviews in Computational Chemistry*, Vol. 5, K. B. Lipkowitz and D. B. Boyd, Eds., VCH, New York, pp. 337-380 (1994); see also, W. C. Guida, *Curr. Opin. Struct. Biology*, 1994, 4:777-781].

Once a compound has been designed or selected by the above methods, the efficiency with which that entity may bind to an human α-GAL binding pocket may be tested and optimized by computational evaluation. For example, an effective human α-GAL binding pocket inhibitor must preferably demonstrate a relatively small difference in energy between its bound and free states (i.e., a small deformation energy of binding). Thus, the most efficient human α-GAL binding pocket inhibitors should preferably be designed with a deformation energy of binding of not greater than about 10 kcal/mole, more preferably, not greater than 7 kcal/mole. Human α-GAL binding pocket inhibitors may interact with the binding pocket in more than one conformation that is similar in overall binding energy. In those cases, the deformation energy of binding is taken to be the difference between the energy of the free entity and the average energy of the conformations observed when the inhibitor binds to the protein.

An entity designed or selected as binding to a human α-GAL binding pocket may be further computationally optimized so that in its bound state it would preferably lack repulsive electrostatic interaction with the target enzyme and with the surrounding water molecules. Such non-complementary electrostatic interactions include repulsive charge-charge, dipole-dipole and charge-dipole interactions.

Specific computer software is available in the art to evaluate compound deformation energy and electrostatic interactions. Examples of software designed for such uses include: Gaussian 94, revision C (M. J. Frisch, Gaussian, Inc., Pittsburgh, PA, ©1995); AMBER, version 4.1 (P. A. Kollman, University of California at San Francisco, ©1995);

-20-

QUANTA/CHARMM (Molecular Simulations, Inc., San Diego, CA, ©1995); Insight II/Discover (Molecular Simulations, Inc., San Diego, CA ©1995); DelPhi (Molecular Simulations, Inc., San Diego, CA ©1995); and AMSOL (Quantum Chemistry Program Exchange, Indiana University). These programs may be implemented, for instance, using a Silicon Graphics workstation such as an INDIGO^2 ™ with IMPACT™ graphics. Other hardware systems and software packages will be known to those skilled in the art.

Another approach enabled by this invention, is the computational screening of small molecule databases for chemical entities or compounds that can bind in whole, or in part, to a human α-GAL binding pocket. In this screening, the quality of fit of such entities to the binding site may be judged either by shape complementarity or by estimated interaction energy (E. C. Meng et al., *J. Comp. Chem.*, 1992, 13:505-524).

According to another embodiment, the invention provides compounds which associate with a human α-GAL-like binding pocket produced or identified by the method set forth above.

The structure coordinates set forth in FIG. 1 can also be used to aid in obtaining structural information about another crystallized molecule or molecular complex. This may be achieved by any of a number of well-known techniques, including molecular replacement.

Therefore, in another aspect this invention provides a method of utilizing molecular replacement to obtain structural information about a molecule or molecular complex whose structure is unknown comprising the steps of:

a) crystallizing said molecule or molecular complex of unknown structure;

b) generating X-ray diffraction data from said crystallized molecule or molecular complex; and

c) applying at least a portion of the structure coordinates set forth in FIG. 1 to the X-ray diffraction data to generate a three-dimensional electron density map of the molecule or molecular complex whose structure is unknown.

By using molecular replacement, all or part of the structure coordinates of the human α-GAL as provided by this invention (and set forth in FIG. 1) can be used to determine the structure of a crystallized molecule or molecular complex whose structure is unknown more quickly and efficiently than attempting to determine such information *ab initio*.

Molecular replacement provides an accurate estimation of the phases for an unknown structure. Phases are a factor in equations used to solve crystal structures that can not be determined directly. obtaining accurate values for the phases, by methods other than

-21-

molecular replacement, is a time-consuming process that involves iterative cycles of approximations and refinements and greatly hinders the solution of crystal structures. However, when the crystal structure of a protein containing at least a homologous portion has been solved, the phases from the known structure provide a satisfactory estimate of the phases for the unknown structure.

Thus, this method involves generating a preliminary model of a molecule or molecular complex whose structure coordinates are unknown, by orienting and positioning the relevant portion of the human α-GAL according to FIG. 1 within the unit cell of the crystal of the unknown molecule or molecular complex so as best to account for the observed X-ray diffraction data of the crystal of the molecule or molecular complex whose structure is unknown. Phases can then be calculated from this model and combined with the observed X-ray diffraction data amplitudes to generate an electron density map of the structure whose coordinates are unknown. This, in turn, can be subjected to any well-known model building and structure refinement techniques to provide a final, accurate structure of the unknown crystallized molecule or molecular complex [E. Lattman, *Meth. Enzymol.*, 1985, 115:55-77; M. G. Rossmann, e d., "The Molecular Replacement Method", *Int. Sci. Rev. Ser.*, No. 13, Gordon & Breach, New York (1972)].

The structure of any portion of any crystallized molecule or molecular complex that is sufficiently homologous to any portion of human α-GAL can be resolved by this method.

In a preferred embodiment, the method of molecular replacement is utilized to obtain structural information about another galactosidase. The structure coordinates of human α-GAL as provided by this invention are particularly useful in solving the structure of other isoforms of α-GAL or other α-GAL-containing complexes.

Furthermore, the structure coordinates of human α-GAL as provided by this invention are useful in solving the structure of α-GAL proteins that have amino acid substitutions, additions and/or deletions (referred to collectively as "human α-GAL mutants", as compared to naturally occurring human α-GAL isoforms. These human α-GAL mutants may optionally be crystallized in co-complex with a chemical entity, such as galactose. The crystal structures of a series of such complexes may then be solved by molecular replacement and compared with that of wild-type human α-GAL. Potential sites for modification within the various binding sites of the enzyme may thus be identified. This information provides an additional tool for determining the most efficient binding interactions, for example, increased hydrophobic interactions, between human α-GAL and a chemical entity or compound.

-22-

The structure coordinates are also particularly useful to solve the structure of crystals of human α-GAL or human α-GAL homologues co-complexed with a variety of chemical entities. This approach enables the determination of the optimal sites for interaction between chemical entities, including between candidate human α-GAL agonists and human α-GAL. For example, high resolution X-ray diffraction data collected from crystals exposed to different types of solvent allows the determination of where each type of solvent molecule resides. Small molecules that bind tightly to those sites can then be designed and synthesized and tested for their human α-GAL agonistic activity.

All of the complexes referred to above may be studied using well-known X-ray diffraction techniques and may be refined versus 1.5-3.5 Å resolution X-ray data to an R value of about 0.20 or less using computer software, such as X-PLOR [Yale University, ©1992, distributed by Molecular Simulations, Inc.; see, e.g., Blundell & Johnson, supra; *Meth. Enzymol.*, vol. 114 & 115; H. W. Wyckoff et al., eds., Academic Press (1985)]. This information may thus be used to optimize known human α-GAL agonists/antagonists/inhibitors, and more importantly, to design new human α-GAL agonists/antagonists/inhibitors.

The invention will be more fully understood by reference to the following examples. These examples, however, are merely intended to illustrate the embodiments of the invention and are not to be construed to limit the scope of the invention.

-23-
Examples

Experimental Procedures
Materials and Methods

5   **Cloning and Expression of human α-Galactosidase:**

Human α-Galactosidase (REPLAGAL™ lot G302-010, Transkaryotic Therapies, Inc.) was produced using gene activation technology as described in detail in U.S. Patents Nos. 5,733,761, 6,270,989, and 6,565,844, all of which are expressly incorporated herein by reference. Briefly, regulatory (e.g., a viral promoter) and structural DNA sequences were

10   inserted upstream of the endogenous human α-Galactosidase genomic locus (GenBank Acc. No. HSU78027) in a human cell (e.g., HT-1080) using homologous recombination. As a result, α-Galactosidase expression was enhanced resulting in secretion of α-Galactosidase protein to the culture supernatant. The α-Galactosidase polypeptide was then highly purified using the methods described in detail in U.S. Patents Nos. 6,083,725, 6,395,884 and

15   6,458,574, all of which are expressly incorporated herein by reference.

Crystallization and x-ray data collection:

Human α-Galactosidase was concentrated to 40mg/ml in 20mM TrisHCl pH 7.5 prior to crystallization trials. Crystals were grown in either hanging or sitting drops via vapor

20   diffusion against a reservoir solution of 30% polyethylene glycol (PEG) 4000 (Fluka), 100mM TrisHCl pH 8.0, and 200mM ammonium sulfate. Crystals were then harvested into 35% PEG 4000, 100mM TrisHCl pH 7.5, and 20% (v:v) ethylene glycol. Crystals were cooled in liquid nitrogen and then transferred into a gaseous nitrogen stream at 100K for x-ray data collection. Ligand-soaked crystals were transferred into 31% PEG 3350, 100mM

25   sodium acetate pH 5.5, and 110mM D-(+)-galactose (Sigma) prior to nitrogen cooling and x-ray data collection. Despite efforts to increase their size, the crystals never grew larger than 30 x 30 x 100 μm. For each crystal, 180° of diffraction data were collected at beamline 22-ID at the Advanced Photon Source. Processing of x-ray images using the HKL2000 package (Otwinowski, Z. & Minor, W., *Methods in Enzymology*, 1997, 276:307-326) revealed unit

30   cell constants of approximately 89Å x 89Å x 215Å in space $P3_121$ or $P3_221$. The diffraction from these crystals proved to be extremely anisotropic, with reflections visible to 2.8Å in the direction of the crystallographic *c* axis, but only to approximately 4Å in the perpendicular directions. This, plus the high redundancy and weak diffraction overall from the small crystals, resulted in very poor merging statistics. The native frames were initially processed

-24-

using HKL2000 to 3.25Å. Reprocessing the frames in MOSFLM and SCALA (Collaborative Computational Project, *Acta Crystallogr.*, 1994, D50:760-763) with anisotropic diffraction limits produced maps of lower quality, so this route was abandoned, and the original data were used throughout the refinement. The high resolution limits were determined from the

5   shell where $<I/\sigma_I>$ dropped to 2. Intensities were adjusted with TRUNCATE (Collaborative Computational Project, *Acta Crystallogr.*, supra) prior to molecular replacement and refinement.

Phasing, model building, and refinement:

10   Molecular replacement calculations were performed in the program AmoRe (Collaborative Computational Project, *Acta Crystallogr.*, 1994, D50:760-763) using a homology model of the human α-GAL protein built from the crystal structure of chicken α-NAGAL (Garman, S. C., et al., *Structure*, 2002, 10:425-434). The dimeric model was rotated and translated against the 8-4Å diffraction amplitudes. Molecular replacement in both

15   enantiomorphic space group possibilities identified a dimer of α-GAL in the asymmetric unit of space group $P3_221$ as the top solution, with a correlation coefficient of 28 and an R_{factor} of 58. Inspection of the packing showed no steric clashes in a unit cell with 50% solvent content. Rigid body refinement in the programs AMoRe and CNS (Brünger, A. T., et al., *Acta Crystallogr. D Biol. Crystallogr.* 1998, 54:905-21) was followed by model building in the

20   program O (Jones, T. A., et al., *Acta Crystallogr.*, 1991, A47:110-9). Residue numbering of the α-GAL protein begins at the secretory signal; the mature protein begins at amino acid 32. Refinement protocols in CNS included conjugate gradient minimization, simulated annealing, and temperature factor refinement. Models were built into σ_A weighted simulated annealing composite omit maps calculated in CNS. Strong two-fold non-crystallographic symmetry

25   restraints (300 kcal/mol-$Å^2$) were imposed on all atoms in the early stages of refinement, and later relaxed for the atoms that differ between the two halves of the dimer, including those in crystal contacts and N-linked carbohydrate atoms. Refinement steps were accepted only if they reduced the R_{free} (of a test set comprised of 820 reflections, 5% of the total, selected using resolution shells). The R_{work} and R_{free} for the native structure are 26.2% and 30.1%,

30   respectively, using all reflections. Because of the limited resolution, side chain rotamers were typically chosen during manual rebuilding to be consistent with the 1.9Å chicken α-NAGAL structure.

-25-

Sequence alignments, calculations, and figures:

Beginning with the human α-GAL sequence, a BLAST search (Altschul, S. F., et al., *Nucleic Acids Res* 1997, 25:3389-402) of the NCBI non-redundant protein sequence database found the 50 closest sequences. After removal of 10 highly redundant sequences, the remaining 40 sequences were multiply aligned in CLUSTALW (Thompson, J. D., et al., *Nucleic Acids Res,* 1994, 22:4673-80), then converted into a phylogeny tree using the programs WEIGHBOR (Bruno, W. J., et al., *Mol Biol Evol,* 2000, 17:189-97) and PHYLIP (Felsenstein, J., Phylogeny Inference Package version 3.6, 1995, Department of Genetics, University of Washington, Seattle, WA). The accession codes of the 40 sequences from the NCBI non-redundant database are: NP_000160, NP_038491, CAC44626, XP_318652, AAM29494, XP_315871, NP_611119, AAL87527, XP_235515, NP_000253, 1KTB, NP_506031, NP_822650, NP_624613, AAC99325, NP821803, BAB83765, ZP_00066516, AAM13199, AAP04002, AAG13536, BAC55816, NP_568193, CAC08337, Q42656, BAC66445, T06388, T10860, P14749, AAF04591, BAB12570, NP191190, S45453, P41947, NP_595012, AAG24511, AAB35252, JC5558, NP_811977, and P28351. Sequence identities were calculated without signal sequences in EMBOSS using a Needleman-Wunsch full path matrix algorithm with the BLOSSUM62 matrix, a gap penalty of 10, and a gap extension penalty of 0.5 (Needleman, S. B. & Wunsch, C. D., *J Mol Biol* 1970, 48:443-53). Least squares superpositions of coordinates were performed using the program LSQMAN (Kleywegt, G. J. & Read, R. J., *Structure,* 1997, 5:1557-1569) with a distance cutoff of 3.8Å, and coordinate transformations were applied using the program MOLEMAN2 (Kleywegt, G. J. & Read, R. J., *Structure,* 1997, 5:1557-1569). Molecular figures were prepared using the programs MOLSCRIPT (Kraulis, P. J., *J. Appl. Crystallogr.,* 1991, 24:946-950), BOBSCRIPT (Esnouf, R. M., *J. Mol. Graph. Model.* 1997, 15:132-34), and GRASP (Nicholls, A., et al., *Proteins* 1991, 11:281-96).

Results

The structure of human α-GAL was determined by x-ray crystallographic methods to a resolution limit of 3.25 Å (see Table 1 below).

Table 1: Crystallographic Statistics

Data		
	Native	Ligand
Beamline	APS 22-ID	APS 22-ID
Wavelength, Å	1.033	1.033

-26-

Space Group	$P3_221$	$P3_221$
Cell Lengths, Å	88.5, 88.5, 215.5	90.0, 90.0, 216.5
Resolution, Å (last shell)	50-3.25 (3.37-3.25)	50-3.45 (3.57-3.45)
No. of Observations (last shell)	156309 (9921)	91651 (8610)
No. of Unique Observations (last shell)	16080 (1542)	13922 (1323)
Completeness, % (last shell)	99.8 (98.7)	99.7 (98.9)
Multiplicity (last shell)	9.7 (6.5)	6.6 (6.7)
R_{sym} (last shell)	0.246 (0.740)	0.200 (0.745)
$<I/\sigma_I>$ (last shell)	9.1 (2.4)	8.2 (2.4)
Refinement		
R_{work} / R_{free}	26.2% / 30.1%	28.5% / 32.1%
No. of Atoms: Protein	6251	6251
Carbohydrate	268	331
Other	18	18
Ramachandran: Favored	74.4%	74.3%
Allowed	23.0%	23.8%
Generous	2.5%	1.5%
Forbidden	0%	0.4%
RMS Deviations: Bonds	0.009 Å	0.008 Å
Angles	1.5°	1.5°
Dihedrals	22.8°	22.8°
Impropers	0.9°	0.8°

$R_{sym} = \Sigma_h \Sigma_i |I_{h,i} - <I_h>|/\Sigma_h \Sigma_i |I_{h,i}|$, where $I_{h,i}$ is the ith intensity measurement of reflection h and $<I_h>$ is the average intensity of that reflection.

$R_{work}/R_{free} = \Sigma_h |F_P - F_C|/\Sigma_h |F_P|$, where F_C is the calculated and F_P is the observed structure factor amplitude of reflection h for the working/free set, respectively.

The x-ray structure reveals human α-GAL as a homodimeric glycoprotein with each monomer composed of two domains, a $(\beta/\alpha)_8$ domain containing the active site and a C-terminal domain containing eight antiparallel β strands on two sheets in a β sandwich. After removal of the 31 residue signal sequence, the first domain extends from residues 32 to 330 and contains the active site formed by the C-terminal ends of the β strands at the center of barrel, a typical location for the active site in $(\beta/\alpha)_8$ domains. The second domain, comprised of residues 331 to 429, packs against the first with an extensive interface, burying 2500 Å^2 of surface area within one monomer. The dimer has overall protein dimensions of approximately 75 x 75 x 50Å. The molecule is concave in the third dimension and varies in thickness from approximately 20 to 50Å. Electron density is visible for 390 and 391 amino acid residues (out of 398 total) in the two copies of the monomer in the crystallographic asymmetric unit; the missing residues occur at the C-terminus. The two monomers pack with

-27-

an interface that extends the 75Å width of the dimer and buries 2200 Å^2 of surface area. In the dimer interface, 30 residues from each monomer contribute to the interface, from loops β1-α1, β6-α6, β7-α7, β8-α8, β11-β12, and β15-β16. The dimer is markedly negatively charged, as seen in a surface electrostatic potential. With 47 carboxylate groups and only 36 basic residues in the 398 residues in the molecule, the overall charge per monomer is expected to be −11 at neutral pH. The carboxylates are most concentrated around the active site, but in the low pH of the lysosome, many of these groups become protonated, reducing the charge on the molecule. In addition to the negative charges on the protein, the N-linked carbohydrate is highly phosphorylated and sialylated (Lee, K., et al., *Glycobiology*, 2003, 13:305-13), further increasing its negative electrostatic potential. The N-linked carbohydrates fall distal to the active sites. Each monomer contains the three N-linked carbohydrate sites, five disulfide bonds (C52-C94, C56-C63, C142-C172, C202-C223, and C378-C382), two unpaired cysteines (C90 and C174), and three *cis* prolines (P210, P380, and P389).

As mentioned above, the C-terminal seven and eight residues of each chain have no electron density associated with them and are presumably disordered. This disorder is consistent with the observation of slight heterogeneity in the C-terminus of recombinant human α-GAL, where the truncation of one or two residues from the C-terminus can occur but has no effect upon the activity of the enzyme (Lee, K., et al., *Glycobiology*, 2003, 13:305-13). The structure offers no support for the observation that the removal of 2 to 10 residues from the C-terminus increases the activity of α-GAL (Miyamura, N., et al., *J Clin Invest*, 1996, 98:1809-17), because the final residue seen in the structure falls at least 45Å from each active site and on the opposite face of the molecule.

Substrate specificity and catalytic mechanism

In both the native and galactose-soaked crystal structures, electron density appears in the two crystallographically-independent active sites. In the galactose-soaked crystal, this density represents α-galactose, the normal catalytic product of the enzyme (K_i ~1mM). In the native structure, this density most likely derives from the cryoprotectant ethylene glycol, a weak inhibitor of glycoside hydrolases (Tsitsanou, K. E., et al., *Protein Sci*, 1999, 8:741-9), analogous to the insertion of glycerol into carbohydrate binding sites on proteins (Garman, S. C., et al., *Structure*, 2002, 10:425-434; Tsitsanou, K. E., et al., *Protein Sci*, 1999, 8:741-9; Schmidt, A., et al., *Protein Sci*, 1998, 7:2081-8). The two active sites of the dimer are separated by approximately 50 Å. As the enzyme shows little change between the liganded

-28-

and unliganded structures, there is no evidence for cooperativity between the two sites, although the biochemical evidence is mixed (Dean, K. J. & Sweeley, C. C., *J Biol Chem*, 1979, 254:9994-10000; Bishop, D. F. & Desnick, R. J., *J Biol Chem*, 1981, 256:1307-16).

We have determined that human α-GAL binds α-galactose by making specific contacts to each functional group on the monosaccharide. Residues from seven loops in domain 1 form the *active site*: β1-α1, β2-α2, β3-α3, β4-α4, β5-α5, β6-α6, and β7-α7. The active site is formed by the side chains of residues W47, D92, D93, Y134, C142, K168, D170, E203, L206, Y207, R227, D231, D266, and M267. Thus, a binding pocket defined by the structural coordinates of these amino acids, as set forth in FIG. 1; or a binding pocket whose root mean square deviation from the structure coordinates of the backbone atoms of these amino acids is not more than 1.5 Å is considered a human α-GAL-like binding pocket of this invention. In important embodiments, C172 makes a disulfide bond to C142.

In the α-GAL/α-NAGAL family, specificity for the 2 position on the galactose ligand occurs via the β5-α5 loop. This was called the "N-acetyl recognition loop" in α-NAGAL (Garman, S. C., et al., *Structure*, 2002, 10:425-434); in the overall α-GAL/α-NAGAL family "2 position recognition loop" or "2 loop" is appropriate. This loop falls near the boundary of exons 4 and 5 of animal α-GAL/α-NAGAL, which have a small insertion in this region, resulting in a short helical stretch at the top of the β5 strand; this insertion is absent in other species. Plant and fungal α-GALs use a Cys and a Trp on this loop to coordinate the 2-hydroxyl on galactose; animal α-GAL uses a Glu and a Leu to recognize the 2-hydroxyl while animal α-NAGAL uses a Ser and an Ala to recognize an N-acetyl at the 2 position. In the animal enzymes, the larger Glu and Leu side chains sterically block the larger N-acetyl substituent, while the smaller Ser and Ala side chains nicely accommodate an N-acetyl group and tolerate a hydroxyl group.

With three different conformations in the 2 loop now identified, the substrate specificity of the other members of the family can be categorized by homology. For example, genome sequencing of *Drosophila melanogaster* and *Anopheles gambiae* have each identified pairs of genes in the α-GAL family. By examination of the sequences in the 2 loop, two are clearly α-NAGALs while the other two appear to be α-GALs. Surprisingly, *Aspergillus niger* contains an enzyme identified as α-GAL that, although only 30% identical to the animal protein sequences, contains a 2 loop virtually identical to animal α-NAGALs. We predict this enzyme is primarily an α-NAGAL with partial α-GAL activity, much like

-29-

human α-NAGAL, which was originally thought to be an α-GAL based upon similar activity (Dean, K. J., et al., *Biochem. Biophys. Res. Commun.*, 1977, 77:1411-7; Schram, A. W., et al., *Biochim. Biophys. Acta,* 1977, 482:138-44).

Although human α-GAL makes contacts to each functional group on the α-galactose ligand, the enzyme shows little specificity for the distal portion of the substrate beyond the glycosidic linkage, and the active site cleft is found in a broad opening on the concave surface of the enzyme. The lack of substrate specificity of human α-GAL beyond the terminal α-galactose differs slightly from the specificity of other α-GALs, which act only upon substrates containing terminal α1-6 galactose groups (Kim,W.D., et al., *Phytochemistry,* 2002, 61:621-30). This increased specificity of plant α-GALs may derive from their monomeric structure, as residues buried in the dimer interface of animal α-GALs (*e.g.,* those on the β1-α1 loop - Fujimoto, Z., et al., *J Biol Chem,* 2003, 278:20313-8) are available for ligand recognition in monomeric α-GALs.

Both α-GALs and α-NAGALs are α retaining exoglycosidases, where both the substrate and product of the catalytic reactions are α anomers at the 1 position on the galactose ring. This retention of anomeric configuration is accomplished by a double displacement catalytic mechanism where the anomeric carbon undergoes two successive nucleophilic attacks (Vasella, A., et al., *Curr Opin Chem Biol,* 2002, 6:619-29). The two sequential inversions of the anomeric carbon lead to retention of the configuration at the end of the catalytic cycle. In two α-GALs from different species, peptic digestion of covalently trapped intermediates has identified the specific aspartic acid acting as the catalytic nucleophile (Hart, D. O., et al., *Biochemistry,* 2000, 39:9826-36; Ly, H. D., et al., *Carbohydr. Res.,* 2000, 329:539-47). These data, combined with the high resolution structure of chicken α-NAGAL, predict the catalytic mechanism of human α-GAL. In human α-GAL, the first nucleophilic attack upon the substrate comes from D170, cleaving the glycosidic linkage and leading to a covalent enzyme-intermediate complex. In the second step of the reaction, a water molecule (deprotonated by D231) attacks C1 of the covalent intermediate, liberating the second half of the catalytic product and regenerating the enzyme in its initial state. Human α-GAL operates most efficiently at low pH, consistent with its highly acidic composition and its lysosomal location.

Retaining glycosidases typically have distances of 5–6Å between catalytic carboxylates, while inverting glycosidases typically have distances of 9–11Å between these residues (McCarter, J. D. & Withers, S. G., *Curr. Opin. Struct. Biol.* 1994, 4:885-92). From

-30-

these distances, it has been possible to reliably predict the mechanism and function of a glycosidase given its structure. However, this rule must be reconsidered in light of the new structures in the α-GAL/α-NAGAL family: for the known structures in the family, the closest approach of the two catalytic carboxylates is 6.5-7Å, among the largest distances seen for retaining glycosidases.

Comparison to related molecules

Human α-GAL is most closely related to α-NAGAL, with the human enzymes sharing 49% amino acid sequence identity. A phylogeny tree of the 40 proteins most closely related to human α-GAL reveals that vertebrate α-GAL and α-NAGAL cluster and have evolved from a common precursor (Wang, A. M., et al., *J. Biol. Chem.*, 1990, 265:21859-66; Wang, A. M., et al., *Mol. Genet. Metab.*, 1998, 65:165-73), while plant and other α-GALs segregate into distinct clusters. The 40 proteins share from 32 to 78% amino acid sequence identity with human α-GAL, with the sequence conservation higher in domain 1, particularly among residues forming the active site.

The 40 sequences include two structures of a family of 27 glycoside hydrolases: human α-GAL and chicken α-NAGAL (Garman, S. C., et al., *Structure*, 2002, 10:425-434) (51% amino acid identity with human α-GAL). Both enzymes share common tertiary structures: each monomer contains both a $(\beta/\alpha)_8$ N-terminal domain and an antiparallel β C-terminal domain. The N-terminal domains superimpose very well: the chicken α-NAGAL superimposes on the human α-GAL with a root mean square deviation (RMSD) of 0.7Å for 295 Cα atoms. Domain 2, with lower sequence conservation, superimposes less well: the chicken domain superimposes on human with an RMSD of 1.3Å for 80 Cα atoms. The most important residue in the dimer interface, F273, has 130Å^2 surface area buried per monomer upon formation of the dimer. This residue alone (out of the 30 in the dimer interface) accounts for 12% of the buried surface area in the interface. This residue is a Phe or Tyr in most animal α-GALs and α-NAGALs, while in plant α-GALs, the equivalent residue is a Gly. Thus, this residue predicts the dimerization state of the enzyme in different species: Phe or Tyr indicates the enzyme is a dimer, while Gly indicates the enzyme remains a monomer.

N-linked carbohydrate and lysosomal targeting

Both endogenous and recombinant α-GAL show a large amount of heterogeneity in the attached carbohydrate, with over 70 different glycoforms (Lee, K., et al., *Glycobiology,*

-31-

2003, 13:305-13; Bishop, D. F. & Desnick, R. J., *J Biol Chem*, 1981, 256:1307-16; Matsuura, F., et al., *Glycobiology*, 1998, 8:329-39; LeDonne, N. C., et al., *Arch Biochem Biophys*, 1983, 224:186-95; Ioannou, Y. A., et al., *Biochem J.*, 1998, 332:789-97). Despite the resolution of the human α-GAL structure, extensive density appears for N-linked carbohydrates. Each
5   monomer has four potential N-linked carbohydrate attachment sites (N139, N192, N215, and N408), the first three of which show carbohydrate electron density. The fourth potential site at N408 contains the amino acid sequence Asn-Pro-Thr, a sequence not ordinarily recognized by the carbohydrate attachment machinery (Gavel, Y. & von Heijne, G., *Protein Eng*. 1990, 3:433-42), consistent with the absence of carbohydrate at this location in recombinant α-GAL
10 expressed in COS cells (Ioannou, Y. A., et al., *Biochem J.*, 1998, 332:789-97), CHO cells and human cells (Lee, K., et al., *Glycobiology*, 2003, 13:305-13). The three sites with attached carbohydrate show density in both independent monomers in the asymmetric unit and in both the native and ligand-soaked crystals.

The glycosylation pattern differs among the structures in the α-GAL/α-NAGAL
15 family. The chicken α-NAGAL and human α-GAL each contain three sites, two of which (N192 and N215 in α-GAL numbering) are in common. These two carbohydrates are attached to helices α4 and α5, away from the active site and from the dimer interface. The N-linked carbohydrate at N215 is necessary but not sufficient for successful secretion of the active enzyme, and the N192 carbohydrate site improves secretion of the active enzyme
20 (Ioannou, Y. A., et al., *Biochem J.*, 1998, 332:789-97). These two sites have a large proportion of oligomannose-containing carbohydrate, while the N139 site contains no oligomannosyl carbohydrate, only complex carbohydrate (Lee, K., et al., *Glycobiology*, 2003, 13:305-13). Thus the N-linked carbohydrate at N192 and N215 is responsible for targeting the glycoprotein to the lysosome, because only oligomannosyl carbohydrates contain the
25 lysosomal targeting signal, mannose-6-phosphate (Ghosh, P., et al., *Nat Rev Mol Cell Biol*, 2003, 4:202-12). The N192 and N215 side chains are 20Å apart on the same face of the molecule, 24 and 23Å away from the active site respectively. Unlike many N-linked carbohydrates that lie along the surface of the protein and shield surface-exposed hydrophobic residues, the carbohydrate at N215 extends away from the protein, in an ideal
30 position to bind to the mannose-6-phosphate receptor (M6PR). Mutation of N215 to Ser eliminates the carbohydrate attachment site, causing inefficient trafficking of the enzyme to the lysosome (Ioannou, Y. A., et al., *Biochem J.*, 1998, 332:789-97) and leading ultimately to the development of Fabry disease (Davies, J. P., et al., *Hum Mol Genet*, 1993, 2:1051-3).

-32-

Unique among the carbohydrate attachment sites, N215 shows different primary glycoforms in the two recombinant enzymes used as Fabry disease treatments: in REPLAGAL™ this site is mostly singly phosphorylated oligomannose, while in FABRAZYME™ this site is mostly biphosphorylated oligomannose (Lee, K., et al., *Glycobiology*, 2003, 13:305-13). The M6PR transport pathway is also used by the recombinant glycoprotein in the treatment for Fabry disease: upon injection into the bloodstream of a Fabry patient, the recombinant glycoprotein is delivered into the lysosomes of affected cells via M6PR on the surface. The pharmacological differences between the REPLAGAL™ and FABRAZYME™ α-GAL preparations derive from the different glycoforms attached to N192 and N215.

Part of the experiments described elsewhere herein are also described in detail in a publication by Garman, S. C. & Garboczi, D. N., entitled "The Molecular Defect Leading to Fabry Disease: Structure of Human α-Galactosidase", *J. Mol. Biol.*, 2004, 337:319-335, the contents of which are expressly incorporated herein by reference.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.

All references disclosed herein are incorporated by reference in their entirety. What is claimed is presented below and is followed by a Sequence Listing.

We claim: